VINF Dokumentácia

Vyhľadávanie pôvodu slova v akomkoľvek jazyku

Richard Andrášik

1. Sumarizácia projektu

Tento program primárne slúži na vyhľadanie pôvodu slov. Jeho hlavná funkcionalita sa skladá z dvoch častí. Najprv je slovo preložené do originálu. Potom je nájdený pôvod hľadaného slova. Ako dataset je použitý enwictionary, slovník od wikipédie s veľkosťou 7,4 GB. Pred spustením programu je potrebné vytvoriť index.csv pomocou testindex.py. Hlavný program použije tento index na jednoduchšie vyhľadanie slov. Súbor testindex.py obsahuje funkcie Sparku, ktoré zrýchľujú indexovanie celého datasetu.

2. Cesta k finálnemu výsledku

Pri prvej konzultácii som mal vymyslené, čo program bude robiť a ako ho budem môcť rozšíriť. Mal som už aj nápady o rozšírenej funkcionalite projektu.

Pre druhú konzultáciu som vytvoril pseudokód na základe ktorého som napísal potom program, ktorý dokáže v datasete nájsť pôvod anglického slova. Pracoval som vtedy iba s malým datasetom, asi 257kB. Program vtedy prehľadával iba tituly stránok a etymológie priamo príslušné ku nim, takže celý dataset bolo potrebné prejsť iba raz. Preto každé spustenie trvalo rovnako veľa času.

V tretej konzultácii som mal základnú funkcionalitu hotovú t.j. nájdenie pôvodu slova pre akékoľvek slovo. Taktiež som vytvoril konzolové UI pre lepšie používanie aplikácie, vysoko som tým zlepšil user experience. Pridal som regexy na efektívnejšie vyhľadávanie prekladov slov. Pri tretej konzultácii som používal aj prvú verziu môjho indexu. Program na indexáciu bol vtedy veľmi jednoduchý, písal do súboru slová, ich jazyky a riadok začiatkov strán oddelených oddeľovačom, na ktorých sa nachádzajú.

Do štvrtej konzultácie som urobil od základu úplné prepracovanie programu. Vytvoril som nové konzolové UI, s lepšou prehľadnosťou a ošetrenými vstupmi. Dokončil som bonusovú funkcionalitu, teda navrhnutie podobný slov po nájdení pôvodu slova a možnosť vyhľadať pôvod nového slova. Od tejto konzultácie môj program už pracuje dostatočne efektívne aby našiel pôvod akéhokoľvek slova v celom datasete len za pár minút.

Pre poslednú konzultáciu som pridal paraelizáciu pomocou knižnice pyspark. Paraelizoval som v indexácii vyhľadávanie prekladov slov, keďže tam program ide po riadkoch, bolo to ideálne miesto, čo sa dá paraelizovať. Kvôli paraelizácii bolo potrebné prerobiť index aby ukazoval na riadky, kde sa slová nachádzajú namiesto riadkov, kde začínajú strany v ktorých sú. Po prerobení indexovania bolo potrebné prerobiť aj hlavný program aby našiel celé strany pomocou týchto riadkov, na ktorých sú.

Po konzultácii som dopísal ešte dokumentáciu a pridal viac komentárov do obidvoch programov.

3. Návod na spustenie

1) Stiahnuť etymology.py a testindex.py a vložiť ich do toho istého priečinku.

2) Stiahnuť verziu enwictionary.xml a zmeniť dva path riadky v obidvoch programoch aby smerovali na stiahnutý enwictionary.

3) Spustiť testindex.py v priestore, ktorý má nainštalovaný pyspark.

4) Spustiť etymology.py a riadiť sa inštrukciami v termináli

4. Testovanie

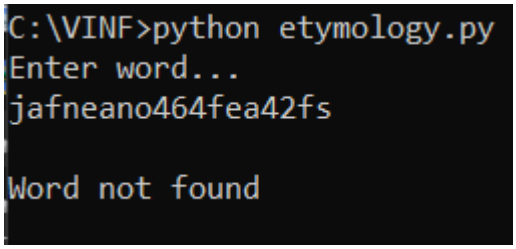
a. Ošetrenie zlých vstupov

Ošetrené vstupy: Zlé slovo zadané, výber nesprávneho čísla pri výbere zo zoznamu, zlý vstup pri výbere z číselného zoznamu

Test 1 : Zadanie nesprávneho slova na vyhľadávanie

Parameter: po prompte zadané slovo, ktoré nie je v nijakom jazyku

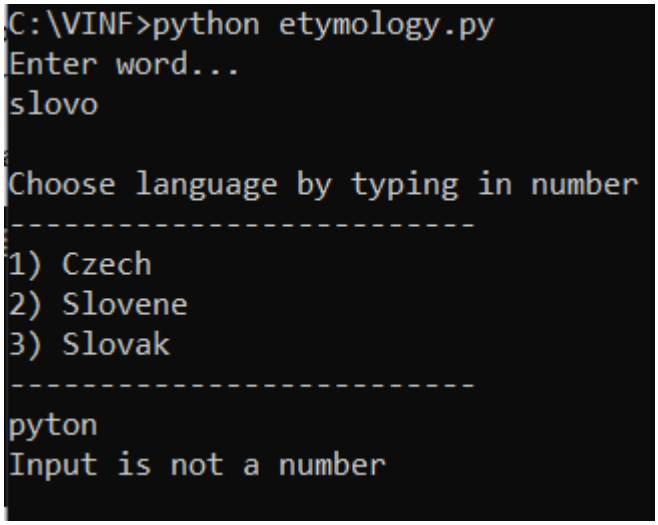
Očakávaný výsledok: program upozorní používateľa, že slovo sa nenašlo



Test 2 : Zadanie písmen pri výbere zo zoznamu

Parameter: po číselnom výbere zadané slovo používateľom

Očakávaný výsledok: program upozorní používateľa o zlom vstupe



Test 3 : Zadanie zlého čísla pri extra funkcionalite

Parameter: po zozname podobných slov zadané číslo, ktoré nie je v zozname

Očakávaný výsledok: program upozorní používateľa, že zadal zlé číslo

```
==========================
Search for similar words?
--------------------------
1: Yes
2: No
--------------------------
1
--------------------------
1: solvo
2: sloop
3: sloot
4: slobo
5: slova
6: ŏlovo
7: salvo
8: stovo
9: úslov
10: silvo
--------------------------
11
Input is out of range
```

Výsledok, všetky 3 testy dopadli úspešne. Testovanie ošetrenia vstupu má 100% úspešnosť. Doplnok, je ošetrených o mnoho viac vstupov ako je tu zobrazených.

b. Spoľahlivosť najdenia správneho slova

Tu budeme vyhľadávať pôvod vybraných 4 slov. Podľa výsledkov uvidíme, či sme dostali hľadané informácie.

Test 1: slovo -> pôvod slova “word” v angličtine

Očakávaný výsledok: možné pôvody slova word v angličtine

```
Enter word...
slovo

Choose language by typing in number
--------------------------
1) Czech
2) Slovene
3) Slovak
--------------------------
3
Found page: word  -  English
==========================
In language English:
[[File:About- General sign.ogv|thumb| The word ''about'' signed in [[American Sign Language]].]]
{{root|en|ine-pro|*werh₁-|*dʰeh₁-}}
From {{inh|en|enm|word}}, from {{inh|en|ang|word}}, from {{inh|en|gmw-pro|*word}}, from {{inh|en|gem-pro|*wurdą}}, from
{{inh|en|ine-pro|*werdʰh₁om|*wr̥dʰh₁om}}. {{doublet|en|verb|verve}}; further related to {{m|en|vrata}}.
--------------------------
In language English:
Variant of {{m|en|worth||to become, turn into, grow, get}}, from {{inh|en|enm|worthen}}, from {{inh|en|ang|weorþan|t=to
turn into, become, grow}}, from {{inh|en|gmw-pro|*werþan}}, from {{inh|en|gem-pro|*werþaną|t=to turn, turn into, become}
}. More at {{section link|worth#Verb}}.
--------------------------
```

Test 2: obrubník [cz] -> pôvod slova “curb“ v angličtine

Očakávaný výsledok: možné pôvody slova curb v angličtine

```
Enter word...
obrubník

Choose language by typing in number
--------------------------
1) Slovak
2) Czech
--------------------------
2

Do you mean?
--------------------------
1) kerb  -  English
2) curb  -  English
--------------------------
2
==========================
In language English:
{{root|en|ine-pro|*(s)ker- (turn)}}
From {{der|en|frm|courbe||curve, curved object}}, from {{der|en|la|curvus||bent, crooked, curved}}. {{doublet|en|curve}}
.
--------------------------
```

Test 3: पंख [lienka v hindi] -> pôvod slova “ladybird“ v angličtine

Očakávaný výsledok: možné pôvody slova ladybird v angličtine

```
Enter word...
□□□

Choose language by typing in number
--------------------------
1) Hindi
2) Marathi
--------------------------
1

Do you mean?
--------------------------
1) fin  -  Translingual
2) ladybird  -  English
3) petal  -  English
--------------------------
2
==========================
In language English:
From {{compound|en|lady|bird}}, the “lady” here referring to the [[Virgin Mary]], Jesus’ mother. Compare {{cog|de|Marien
käfer|lit=Mary beetle/bug}}.
--------------------------
==========================
```

Dodatok: nemám hindi language pack na mojom počítači, preto sú tam znaky otázniky

Test 4: chémia -> pôvod slova “chemistry” v angličtine

Očakávaný výsledok: možné pôvody slova chemistry v angličtine

```
Enter word...
chémia

Choose language by typing in number
--------------------------
1) Slovak
--------------------------
1
Found page: chemistry  -  English
==========================
In language English:
{{root|en|ine-pro|*ǵʰew-}}
First coined 1605, from {{suf|en|chemist|ry}}. From {{m|en|chemist}}, {{m|en|chymist}}, from {{derived|en|la|alchimista}
}, from {{derived|en|ar|□□□□□□|□□□□□□□□□□□□□}}, from article {{m|ar|□□□}} + {{derived|en|grc|χυμεία||art of alloying met
als|sc=polytonic}}, from {{m|grc|χύμα||fluid|sc=polytonic}}, from {{m|grc|χυμός||juice|sc=polytonic}}, from {{m|grc|χέω|
|I pour|sc=polytonic}}.
--------------------------
==========================
```

Záver: Všetky testy boli úspešné.

Precision -> P = (4 ∩ 4)/4 = 1 -> 100%

Recall -> R = (4 ∩ 4)/4 = 1 -> 100%

Iného testovania nie je veľmi čo urobiť, lebo program vráti presne etymológiu pre slovo, ktoré používateľ potrebuje bez akejkoľvek odchylky.